A Comparative Study of Association Rule Algorithms for Investment in Related Sector of Stock Market

Rajeev Kumar
M. Tech. (CS)
Department of Computer Science
H. P. University, Shimla, India

Arvind Kalia, PhD.
Professor
Department of Computer Science
H. P. University, Shimla, India

ABSTRACT
Investment in the related stocks in share market plays vital role for investors. Variation in stock price is the barometer of growth of companies/sectors. Association Rule mining is one of the fundamental research topics in data mining and knowledge discovery that identifies interesting relationships between itemsets and predicted the associative and correlative behaviour for new data. In the present study the data of different stocks from National Stock Exchange of India Limited has taken and tried to find out the related stocks through Weka 3.6.5 data mining tool. In this paper four association rule algorithms: Apriori Association Rule, Predictive Apriori Association Rule, Tertius Association Rule and Filtered Associator were considered and the results of these four algorithms presented at different support and confidence level. It was found that Apriori Association Rule provided better results than other algorithms for selection of related stocks for investment in share market.

KEY WORDS
Weka, Association Rule mining, Confidence level, Support level.

INTRODUCION
Data mining deals with the discovery of hidden knowledge, unexpected patterns and new rules from large databases [1]. Data mining is often defined as finding hidden information in a database. Alternatively it has been called exploratory data analysis, data driven discovery and deductive learning [2]. Data mining essentially provides pattern-based retrieval, in which a pattern in the data is first discovered, and then that pattern is used to present information (the pattern itself or outlier data, perhaps) to the user [3].

To acquire the knowledge, data mining discipline offers a useful strategy called Association Rule Mining (ARM). The significance of the rules generated is dictated by the pre defined minimum level of the parameter support, and the generation of rules deemed useful is guided by the pre defined minimum level of the parameter confidence [4].

The National Stock Exchange (NSE) is India's leading stock exchange covering various cities and towns across the country. NSE was set up by leading institutions to provide a modern, fully automated screen-based trading system with national reach. The Exchange has brought about unparalleled transparency, speed & efficiency, safety and market integrity [14]. We downloaded available data for different stocks historical data using NSE website.

In the present study, 11 stocks from steel, cement, infra structure, paint i.e. 11 attributes, 1000 instances from September 2008 to September 2012 last traded price have been taken. We considered these four sectors for collection of data. We have taken very relevant stocks from each sector and stored it into ms excel sheets. After collecting the data from database, we preprocess the data. Preprocessing means that we delete those columns from database having very low count. We have considered last traded price of each stock on trading date then compared it to previous trading date and marked it 'low' if previous date last traded price is higher than the last traded price of traded date, otherwise marked as 'high'. After preprocessing of data, next step is to find the result using the open source data mining tool, Weka 3.6.5.

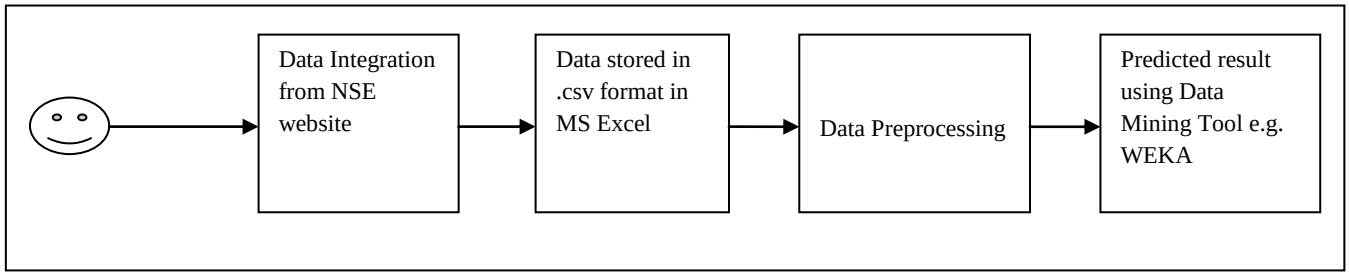

Figure 1: Framework to find the best combination of related stocks in Share Market [5]

Weka is a collection of machine learning algorithms for data mining tasks. The algorithms can either be applied directly to a dataset or called from your own Java code [15]. The Weka workbench contains a collection of visualization tools and algorithms for data analysis and predictive modeling, together with graphical user interfaces for easy access to this functionality. We find out result using four association rule algorithms i.e. Apriori Association Rule, Predictive Apriori Association Rule, Tertius Association Rule and Filtered Associator & compare it. This process is shown in figure1 [5].

LITERATURE REVIEW

Researcher's aim is to compare four association rule algorithms: Apriori Association Rule, Predictive Apriori Association Rule, Tertuis Association Rule and Filtered Associator for course recommender system in E-learning and interpreted according to their simulation result Apriori association algorithm perform better than other algorithms [5].

Dennis P. Groth, Edward L. Robertson, "Discovering Frequent Itemsets", their paper presents new techniques for focusing the discovery of frequent itemsets within large, dense datasets containing highly frequent items. The existence of highly frequent items adds significantly to the cost of computing the complete set of frequent itemsets. Our approach allows for the exclusion of such items during the candidate generation phase of the Apriori algorithm. Afterwards, the highly frequent items can be reintroduced, via an inferencing framework, providing for a capability to generate frequent itemsets without counting their frequency. They demonstrated the use of these new techniques within the well-studied framework of the Apriori algorithm. Furthermore, they provided empirical results using our techniques on both synthetic and real datasets - both relevant since the real datasets exhibit statistical characteristics different from the probabilistic assumptions behind the synthetic data. The source they used for real data was the U.S. Census [6].

Researcher in their paper "Mining Association Rules Between Sets of Items in Large Datasets" stated that they are given a large database of customer transactions. They presented an efficient algorithm that generates all significant association rules between items in database [7].

Researcher stated that the internet is filled with opportunities for learning, communicating, and sharing information and explored the relationship between parent's information literacy, the confidence in child's ability of self-defense on the internet, and adequate measures to promote child using the internet more effectively [8].

In the paper "Comparison and Analysis of algorithm for Association Rules", researcher stated that among AIS algorithm, SETM algorithm, Apriori algorithm, and concluded that SETM algorithm is most efficient also most convenient one to combine DBMS. They also investigated the weakness and strengths of these algorithms [9].

Zijian Zheng, Ron Kohavi, Llew Mason; "Real world performance of association rule algorithms" stated in their study that compares five well-known association rule algorithms using three real-world datasets and an artificial dataset. The experimental results confirm the performance improvements previously claimed by the authors on the artificial data, but some of these gains do not carry over to the real datasets, indicating over fitting of the algorithms to the IBM artificial dataset. More importantly, they found that the choice of algorithm only matters at support levels that generate more rules than would be useful in practice. For support levels that generate less than 1,000,000 rules, which are much more than humans can handle and is sufficient for prediction purposes where data is loaded into RAM, Apriori finishes processing in less than 10 minutes. On their datasets, they observed super-exponential growth in the number of rules. On one of their datasets, a 0.02% change in the support increased the number of rules from less than a million to over a billion, implying that outside a very narrow range of support values, the choice of algorithm is irrelevant [10].

3. ASSOCITION RULE ALGORITHMS

Frequent itemset mining leads to the discovery of association and correlations among items in large transactional or relational datasets [11]. Association rules are used to find the frequent pattern, association or correlation in transaction database. Association rule mining can be used in Basket Data Analysis, Educational Data Mining, Classification, Clustering etc. Association Rule algorithms are Apriori, Sampling, Partitioning & Parallel Algorithm. This section describes the Apriori Association Rule, Predictive Apriori Association Rule, Tertius Association Rule & Filtered Associator algorithm briefly.

3.1 Apriori Association Rule

Apriori Association rule is used to harness the frequent patterns in database. Support & confidence are the normal methods used to measure the quality of association rule.

- **Support** for the association rule X->Y is the percentage of transaction in the database that contains XUY.
- **Confidence** for the association rule is X->Y is the ratio of the number of transaction that contains XUY to the number of transaction that contain X.

Terms related to this algorithm are as follow:

- **Frequent Itemsets:** The set of item which has minimum support & it is denoted by Li for ith itemset.
- **Apriori Property:** Any subset of frequent itemset must be frequent.
- **Join Operation:** To find Lk, a set of candidate k-itemsets is generated by joining Lk-1 with itself.
- **Join Step:** Candidate item Ck is generated by joining Lk-1with itself
- **Prune Step:** Any (k-1)-itemset that is not frequent cannot be a subset of a frequent k-itemset [5].

The Apriori association algorithm is given below [12]:

Algorithm: Apriori Association Rule Algorithm

Purpose: To find subsets which are common to at least a minimum number C (Confidence Threshold) of the itemsets.

Input: Database of Transactions D= {t1, t2, …, tn}
Set if Items I= {I1, I2,…., Ik}
Frequent (Large) Itemset L Support, Confidence.

Output: Association Rule satisfying Support & Confidence

Method:

1. C1 = Itemsets of size one in I;
2. Determine all large itemsets of size 1, L1;
3. i = 1;
4. Repeat
5. i = i + 1;
6. Ci = Apriori-Gen(Li-1);
7. Apriori-Gen (Li-1)
    1. Generate candidates of size i+1 from large

itemsets of size i.
2. Join large itemsets of size i if they agree on
i-1.
3. Prune candidates who have subsets that are
not large.
8. Count Ci to determine Li;
9. until no more large itemsets found;

Figure 2 shows the generation of itemsets & frequent itemsets where the minimum support count is 2.

To generate the association rule from frequent itemset we use the following rule:

- For each frequent itemset L, find all nonempty subset of L
- For each nonempty subset of L, write the association rule S. (L-S) if support count of L/support count of S >= Minimum Confidence

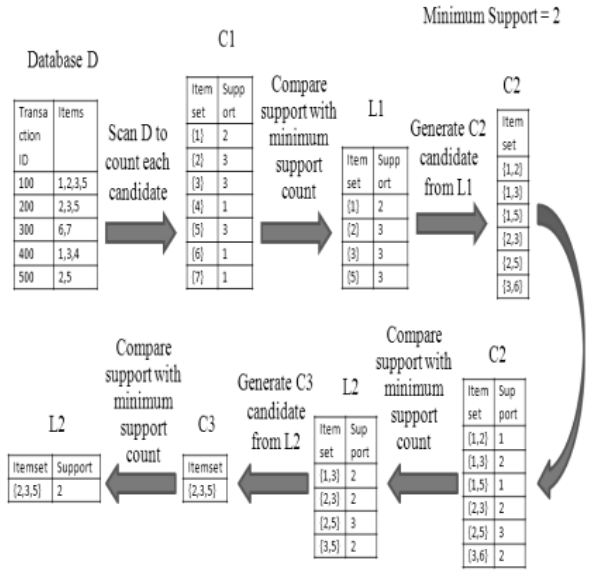

Figure 2: Generation of itemsets & frequent itemsets [5]

The best rule from the itemset L= {2, 3, 5} are calculated as follows:

Consider the minimum support is 2 & minimum confidence is 70%. All nonempty subset of {2, 3, and 5}are: {2,3},{2,5},{3,5},{2},{3},{5}.

Rule 1: {2, 3}. {5} Confidence = Support Count of ({2, 3, 5})/ Support Count of ({2, 3}) = 2/2 = 100%
Rule 2: {2, 5}. {3} Confidence = Support Count of ({2, 3, 5})/ Support Count of ({2, 5}) = 2/3 = 67%
Rule 3: {3, 5}. {2} Confidence = Support Count of ({2, 3, 5})/ Support Count of ({3, 5}) = 2/2 = 100%
Rule 4: {2}. {3, 5} Confidence = Support Count of ({2, 3, 5})/ Support Count of ({2}) = 2/3 = 67%
Rule 5: {3}. {2, 5} Confidence = Support Count of ({2, 3, 5})/ Support Count of ({3}) = 2/3 = 67%
Rule 6: {5}. {2, 3} Confidence = Support Count of ({2, 3, 5})/ Support Count of ({5}) = 2/3 = 67%

Hence the accepted rules are Rule 1 & Rule 3 as the confidence of these rules is greater than 70% [5].

In this paper study, we use Apriori Association algorithm to find out the output i.e. best combination of related stocks, after the preprocessing of data. In Weka the option available with Apriori association rule algorithm are car, class Index, delta, lower bound minimum support, metric type, minimum metric, number of rules, output itemsets, remove all missing columns, significance level, upper bound minimum support, verbose.

3.2 Predictive Apriori Association Rule

In this algorithm, support & confidence is unified into predictive accuracy. This predictive accuracy is used to generate the Apriori association rule. In Weka, this algorithm generates 'n' best association rule based on n selected by the user. It includes options such as car, class index and number of rules to get the result [5].

3.3 Tertius Association Rule

This algorithm finds the rule according to the confirmation measure (P. A. Flach, N. Lachiche 1999). It uses first order logic representation. It allows the user to choose the most convenient or the most comprehensible representation among several possible representations [13]. In Weka it includes various option like class Index, classification, confirmation Threshold, confirmation Values, frequency Threshold, horn Clauses, missing Values, negation, noise Threshold, number Literals, repeat Literals, roc Analysis, values Output etc.

3.4 Filtered Associator

This algorithm is a class for running an arbitrary associator on data that has been passed through an arbitrary filter. Like the associator, the structure of the filter is based exclusively on the training data and test instances will be processed by the filter without changing their structure [5]. In Weka it includes option such as associator with which we can consider the Apriori, Predictive Apriori, Tertius association rule and Filtered Associator algorithm, class index and filter to get the result.

4. EXPERIMENTAL RESULT

The results presented using these four association rule i.e. Apriori Association Rule, Predictive Apriori Association Rule, Tertius Association Rule and Filtered Association Rule shown in table 1.

From Table1 it is predicted in Apriori Association rule that when infrastructure sector remains low/high implies that steel sector also remains low/high following the same for cement and steel sector. This means infrastructure, steel and cement sectors are related in stock market at certain support and confidence level.

In Predicted Apriori Association Rule, it is also predicted that cement, steel and infrastructure sector are related with each other. Here other stocks like paint sector is not related to other stocks.

In Tertuis Association Rule, it is also predicted that steel and infrastructure sector are related but cement and paint sector are not considered in rules generated.
In Filtered associator algorithm, it is predicted that steel, infrastructure and cement sectors are related in stock market but not fast as Apriori algorit

Table1. Result of various association rule algorithm using open source data mining tool WEKA

Stocks Considered	**Association Rule Algorithm using data mining open source tool WEKA**
1.Apriori Association Rule	
1.ACC 2.TSTEEL 3.AMBUJA 4.JSPL 5.JKCEMENT 6.SAIL 7.ASIAN 8.NEROLAC 9.RELINFRA 10.UNITECH 11.LT	Apriori -N 10 -T 0 -C 0.8 -D 0.05 -U 1.0 -M 0.2 -S -1.0 -c -1 **Minimum support: 0.25 (250 instances)** **Minimum metric <confidence>: 0.8** **Number of cycles performed: 15** Best rules found: 1. SAIL=low LT=low 358 ==> TSTEEL=low 307 conf:(0.86) 2. JSPL=low SAIL=low 350 ==> TSTEEL=low 297 conf:(0.85) 3. SAIL=low RELINFRA=low 346 ==> TSTEEL=low 293 conf:(0.85) 4. TSTEEL=low LT=low 363 ==> SAIL=low 307 conf:(0.85) 5. TSTEEL=low UNITECH=low 335 ==> SAIL=low 283 conf:(0.84) 6. ACC=low SAIL=low 320 ==> TSTEEL=low 270 conf:(0.84) 7. SAIL=low UNITECH=low 336 ==> TSTEEL=low 283 conf:(0.84) 8. TSTEEL=low RELINFRA=low 349 ==> SAIL=low 293 conf:(0.84) 9. TSTEEL=low JSPL=low 354 ==> SAIL=low 297 conf:(0.84) 10. JSPL=high SAIL=high 350 ==> TSTEEL=high 293 conf:(0.84) 11. SAIL=high LT=high 356 ==> TSTEEL=high 298 conf:(0.84) 12. JSPL=high LT=high 337 ==> TSTEEL=high 282 conf:(0.84) 13. SAIL=high RELINFRA=high 354 ==> TSTEEL=high 296 conf:(0.84) 14. TSTEEL=high UNITECH=high 357 ==> SAIL=high 297 conf:(0.83) 15. TSTEEL=high RELINFRA=high 356 ==> SAIL=high 296 conf:(0.83) 16. RELINFRA=low LT=low 342 ==> TSTEEL=low 284 conf:(0.83) 17. TSTEEL=low AMBUJA=low 324 ==> SAIL=low 269 conf:(0.83) 18. TSTEEL=high JSPL=high 353 ==> SAIL=high 293 conf:(0.83) 19. TSTEEL=high LT=high 360 ==> SAIL=high 298 conf:(0.83) 20. AMBUJA=low SAIL=low 325 ==> TSTEEL=low 269 conf:(0.83) **Minimum support: 0.3 (300 instances)** **Minimum metric <confidence>: 0.8** **Number of cycles performed: 14** Best rules found: 1. SAIL=low LT=low 358 ==> TSTEEL=low 307 conf:(0.86) 2. TSTEEL=low LT=low 363 ==> SAIL=low 307 conf:(0.85)
2. **Predictive Apriori Association Rule**	
1.ACC 2.TSTEEL 3.AMBUJA 4.JSPL 5.JKCEMENT 6.SAIL 7.ASIAN 8.NEROLAC 9.RELINFRA 10.UNITECH 11.LT	Predictive Apriori -N 10 -c -1 Best rules found: 1. ACC=low JSPL=low JKCEMENT=low NEROLAC=high RELINFRA=low UNITECH=low 60 ==> LT=low 60 acc:(0.99495) 2. TSTEEL=low AMBUJA=low JSPL=low JKCEMENT=low NEROLAC=high RELINFRA=low 58 ==> LT=low 58 acc:(0.99494) 3. AMBUJA=low JSPL=low JKCEMENT=low SAIL=low NEROLAC=high RELINFRA=low 57 ==> LT=low 57 acc:(0.99493) 4. AMBUJA=low JSPL=low JKCEMENT=low NEROLAC=high RELINFRA=low UNITECH=low 55 ==> LT=low 55 acc:(0.99492) 5. ACC=low JSPL=low ASIAN=low NEROLAC=high RELINFRA=low UNITECH=low 53 ==> LT=low 53 acc:(0.99491) 6. AMBUJA=low JSPL=low ASIAN=low NEROLAC=high RELINFRA=low UNITECH=low 48 ==> LT=low 48 acc:(0.99485) 7. ACC=low JSPL=low JKCEMENT=high RELINFRA=low UNITECH=low LT=low 43 ==> TSTEEL=low 43 acc:(0.99475) 8. JSPL=low JKCEMENT=low NEROLAC=high RELINFRA=low UNITECH=low 73 ==> LT=low 72 acc:(0.99472) 9. ACC=low TSTEEL=low AMBUJA=low JSPL=low NEROLAC=high RELINFRA=low UNITECH=low 65 ==> LT=low 64 acc:(0.99445) 10. ACC=low TSTEEL=low ASIAN=high NEROLAC=low UNITECH=low 35 ==> SAIL=low 35 acc:(0.99439)

3. **Tertius Association Rule**	
1.ACC 2.TSTEEL 3.AMBUJA 4.JSPL 5.JKCEMENT 6.SAIL 7.ASIAN 8.NEROLAC 9.RELINFRA 10.UNITECH 11.LT	Tertius -K 10 -F 0.0 -N 1.0 -L 4 -G 0 -c 0 -I 0 -P 0 1. /* 0.545993 0.113000 */ SAIL = low ==> TSTEEL = low 2. /* 0.545993 0.114000 */ TSTEEL = low ==> SAIL = low 3. /* 0.521756 0.051000 */ SAIL = low and LT = low ==> TSTEEL = low 4. /* 0.511881 0.056000 */ TSTEEL = low and LT = low ==> SAIL = low 5. /* 0.498782 0.053000 */ JSPL = low and SAIL = low ==> TSTEEL = low 6. /* 0.494658 0.058000 */ TSTEEL = low ==> SAIL = low or LT = low 7. /* 0.491400 0.053000 */ SAIL = low and RELINFRA = low ==> TSTEEL = low 8. /* 0.491063 0.057000 */ TSTEEL = low and JSPL = low ==> SAIL = low 9. /* 0.490916 0.058000 */ TSTEEL = low ==> SAIL = low or RELINFRA = low 10. /* 0.487543 0.057000 */ TSTEEL = low ==> JSPL = low or SAIL = low Number of hypotheses considered: 23692 Number of hypotheses explored: 3601
4. **Filtered Associator**	
1.ACC 2.TSTEEL 3.AMBUJA 4.JSPL 5.JKCEMENT 6.SAIL 7.ASIAN 8.NEROLAC 9.RELINFRA 10.UNITECH 11.LT	Filtered Associator using weka.associations.Apriori -N 10 -T 0 -C 0.8 -D 0.05 -U 1.0 -M 0.2 -S -1.0 -c -1 on data filtered through weka.filters.MultiFilter –F "weka.filters.unsupervised.attribute.ReplaceMissingValues " Minimum support: 0.25 (250 instances) Minimum metric <confidence>: 0.8 Number of cycles performed: 15 Best rules found: 1. SAIL=low LT=low 358 ==> TSTEEL=low 307 conf:(0.86) 2. JSPL=low SAIL=low 350 ==> TSTEEL=low 297 conf:(0.85) 3. SAIL=low RELINFRA=low 346 ==> TSTEEL=low 293 conf:(0.85) 4. TSTEEL=low LT=low 363 ==> SAIL=low 307 conf:(0.85) 5. TSTEEL=low UNITECH=low 335 ==> SAIL=low 283 conf:(0.84) 6. ACC=low SAIL=low 320 ==> TSTEEL=low 270 conf:(0.84) 7. SAIL=low UNITECH=low 336 ==> TSTEEL=low 283 conf:(0.84) 8. TSTEEL=low RELINFRA=low 349 ==> SAIL=low 293 conf:(0.84) 9. TSTEEL=low JSPL=low 354 ==> SAIL=low 297 conf:(0.84) 10. JSPL=high SAIL=high 350 ==> TSTEEL=high 293 conf:(0.84)

5. CONCLUSION AND FUTURE SCOPE

In this paper study, it is found that Apriori Association Rule algorithm predicted better and fast results in comparison to other algorithms. As seen in the Apriori algorithm results, infrastructure and steel sector are most associated followed by cement sector and its result matched the general real world interdependencies whereas other algorithms were not able to match the real world behaviour. So investor could maximize his profit by investing in related stocks of share market. He could book handsome profit by timely investing and exiting from these associated keen sectors. Future scope can be done with metal, commodities, bank, and information technology sector by using Apriori Association Rule algorithm for related stocks in these sectors.

6. REFERENCES

[1] Pang Ning Tan, Michael Steinbach, Vipin Kumar, 2009. Introduction to Data Mining, Pearson Education, pp. 223.

[2] Li Xiaohong, Huang Jingwei, "SHC: a spectral algorithm for hierarchical clustering", International Conference on "Multimedia Information Networking and Security", 2009.

[3] David W. Cheung, H.Y. Hwang ADA W FU, Jiawei Han, "Efficient Rule-based Attribute-oriented Induction for Data Mining", Journal of intelligent information system, 15,175-200,2000.

[4] A B M Shawkat Ali, Saleh A. Wasimi, 2009. Data Mining: Methods and Techniques, Cengage Learning, 2009 pp(172,175).

[5] Sunita B Aher and Lobo L.M.R.J, "A Comparative Study of Association Rule Algorithms for Course Recommender System in E-Learning" in International Journal of Computer Applications (0975-8887), Volume 39 – No. 1, February2012.

[6] Dennis P. Groth, Edward L. Robertson, "Discovering Frequent Itemsets in the Presence of Highly Frequent Items", Computer Science, Indiana University, Bloomington, IN 47405, USA.

[7] Agrawal Rakesh, Imienski Tmasz; Swami Arun, "Mining Association Rules Between Sets of Items in Large Databases", SIGMOD Conference: 207-216, 1993.

[8] Chia-Chia Lin, Dong-Her Shih, "Associating Information Literacy with Regulating Rules in Family by Data Mining", The 3rd International Conference on Innovative Computing Information and Control, IEEE-2008.

[9] He Lijun Li, Linghua Li, Xiaoniu Wang Degao, "Comparison and Analysis of algorithms for Association

Rules", First International Workshop on Database Technology and Applications, 2009.

[10] Zijian Zheng, Ron Kohavi, Llew Mason, "Real world performance of association rule algorithms", 2001.

[11] Jiawei Han, Micheline Kamber, Jian Pei, 2012. Data Mining Concepts and Techniques, Morgan Kauffman Publishers, pp(244).

[12] Margaret H. Dunham, 2002. Data Mining Introductory and Advanced Topics, Prentice Hall.

[13] P. A. Flach, N. Lachiche, "Confirmation-Guided Discovery of first-order rules with Tertius". Machine Learning, 42:61-95, 2001

[14] http://www.nseindia.com/global/content/about_us/about_us.htm accessed on 20-09-2012

[15] http://www.cs.waikato.ac.nz/~ml/weka/ accessed on 21-09-2012.